DOCUMENT RESUME

ED 401 365 UD 031 361

AUTHOR	Kober, Nancy L., Ed.; Feuer, Michael J., Ed.
TITLE	Title I Testing and Assessment. Challenging Standards for Disadvantaged Children. Summary of a Workshop (Washington, DC, April 1, 1995). Board Bulletin.
INSTITUTION	National Academy of Sciences - National Research Council, Washington, DC. Board on Testing and Assessment.
SPONS AGENCY	Department of Labor, Washington, D.C.
PUB DATE	96
NOTE	31p.
AVAILABLE FROM	Board on Testing and Assessment, National Research Council, 2101 Constitution Avenue, N.W., Washington, DC 20418.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Academic Achievement; *Accountability; *Compensatory Education; *Disadvantaged Youth; Educational Assessment; Educational Change; *Educational Testing; Elementary Secondary Education; Limited English Speaking; State Programs; *Test Construction; Test Use; *Urban Schools; Workshops
IDENTIFIERS	Curriculum Alignment; *Improving Americas Schools Act 1994 Title I

ABSTRACT

The Improving America's Schools Act of 1994 substantially revised the Title I program for disadvantaged children. Some of the most far-reaching changes relate to testing and accountability. States will now need to develop their own assessments and ensure that these assessments are aligned with challenging state standards linked to educational reforms for all students. By the year 2000-2001 states must develop or adopt yearly assessments in reading and mathematics that are aligned with their content and performance standards. To help states in this process, the Board on Testing and Assessment held a half-day workshop on Title I testing and assessment. Mary Jean LeTendre, the Title I director of the Department of Education, opened the workshop by listing assessment issues that included transition policies, the assessment of young children, ensuring adequate progress, measurement problems, multiple measures, the special problems of limited-English- speaking students, and the problems of analyzing disaggregated data. Other speakers discussed:(1) the critical issues facing states as they develop assessments; (2) measurement issues at the state level; (3) the perspectives of test publishers; (4) accountability issues; and (5) issues for further consideration. An appendix gives the agenda and lists participants. (Contains seven references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 401 365

BOARD BULLETIN

Title I Testing and Assessment

Challenging Standards for Disadvantaged Children

Summary of a Workshop

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

K. Barranco

Nat'l Academy Press

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

UD031361

NATIONAL RESEARCH COUNCIL

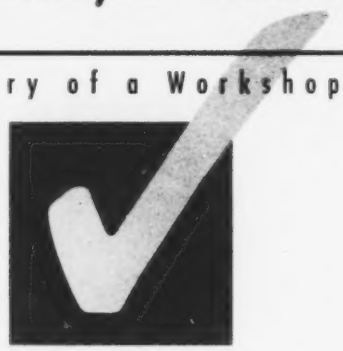
BEST COPY AVAILABLE

BOARD BULLETIN

Title I Testing and Assessment

Challenging Standards for Disadvantaged Children

Summary of a Workshop



Nancy L. Kober and Michael J. Feuer, editors

Board on Testing and Assessment
Commission on Behavioral and Social Sciences and Education
National Research Council

National Academy Press
Washington, D.C. 1996

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Harold Liebowitz is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Kenneth I. Shine is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. Harold Liebowitz are chairman and vice chairman, respectively, of the National Research Council.

The work of the Board on Testing and Assessment is supported by the U.S. Departments of Defense, Education, and Labor, through a grant administered by the U.S. Department of Labor.

Additional copies of this report are available from:
Board on Testing and Assessment
National Research Council
2101 Constitution Avenue N.W.
Washington, D.C. 20418

Copyright 1996 by the National Academy of Sciences. All rights reserved.
Printed in the United States of America

BOARD ON TESTING AND ASSESSMENT
1993-1995

Richard C. Atkinson (*Chair*), University of California, San Diego

Constance B. Newman (*Vice Chair*), Smithsonian Institution, Washington, D.C.

Richard J. Shavelson (*Vice Chair*), School of Education, Stanford University

Laurie J. Bassi, Graduate Public Policy Program, Georgetown University

David C. Berliner, College of Education, Arizona State University, Tempe

Richard F. Elmore, Graduate School of Education, Harvard University

Patricia M. Flynn, Graduate School of Business, Bentley College

Edmund W. Gordon, Department of Psychology, City University of New York

Sylvia T. Johnson, School of Education, Howard University

Brigitte Jordan, Xerox Palo Alto Research Center and Institute for Research on Learning, Palo Alto, Calif.

Carl F. Kaestle, Department of Education, University of Chicago

Michael W. Kirst, School of Education, Stanford University

Luis M. Laosa, Educational Testing Service, Princeton, N.J.

Renée S. Lerche, Ford Motor Company, Dearborn, Mich.

Alan M. Lesgold, Learning Research and Development Center, University of Pittsburgh

Robert L. Linn, School of Education, University of Colorado, Boulder

Miles A. Myers, National Council of Teachers of English, Urbana, Ill.

James L. Outtz, Outtz and Associates, Washington, D.C.

Neal W. Schmitt, Department of Psychology, Michigan State University, East Lansing

Alan H. Schoenfeld, School of Education, University of California, Berkeley

Peter Smith, School of Education and Human Development, George Washington University

David S. Tatel, Hogan and Hartson, Washington, D.C.

William L. Taylor, Attorney at Law, Washington, D.C.

Ewart A.C. Thomas, Department of Psychology, Stanford University

Michael J. Feuer, *Director*

Holly Wells, *Administrative Assistant*

PUBLICATIONS
Board on Testing and Assessment

Anticipating Goals 2000: Standards, Assessment, and Public Policy
Report of a Workshop

Evaluation of the U.S. Employment Service Workplan for the GATB Improvement Project

A Valedictory: Reflections on 60 Years in Educational Testing
By Lee J. Cronbach. Includes the report of the select committee on the California Learning Assessment System

CONTENTS

Preface	vii
Framing Questions	2
State Practitioner Perspectives	5
Measurement Issues	7
Test Publisher Perspectives	10
Accountability Issues	11
Issues for Further Consideration	14
References	16
Appendix: Workshop Agenda and Participants	17

PREFACE

The Improving America's Schools Act of 1994 (P.L. 103-382) substantially revised the Title I program for disadvantaged children, which has been the federal government's most significant effort on behalf of elementary and secondary school students—roughly $7 billion per year—since passage of the original Elementary and Secondary Education Act of 1965. Some of the most far-reaching changes relate to testing and accountability requirements: gone are the federal directives for school districts to use certain kinds of norm-referenced tests or report student test scores in specified formats (see Office of Technology Assessment, 1992). Instead, states will need to develop their own assessments for Title I and ensure that they are aligned with challenging state standards for content and performance linked to state reforms affecting all students. These new assessments will form the basis of a revised Title I accountability system that monitors progress of students toward high learning standards and identifies schools and districts in need of improvement.

The new law took effect July 1, 1995, but states have until school year 1997-98 to set content and performance standards, and until 2000-2001 to adopt new systems of assessment. States vary widely in their measurement expertise, test development resources, and capacity to initiate and implement major new assessment programs; they are therefore at various stages in the process of change. Some have standards, developed through the *Goals 2000: Educate America Act* or another process, but most do not yet have assessments aligned with standards (see also Feuer and Kober, 1995).

Creating a new Title I testing system is one of the most demanding aspects of the new law. Because of its size and scope, changes in the program's testing policies are likely to have significant effects not only on disadvantaged students in American schools, but also on the instruction and assessment of *all* students. State and federal policy makers must confront many complex technical and practical considerations, not the least of which is building an accountability system around a set of assessments that, for the most part, do not yet exist.

It is crucial that states and school districts make wise decisions, based on technically sound information and knowledgeable, practical advice. To help in this process, the Board on Testing and Assessment held a half-day workshop on Title I testing and assessment (on April 1, 1995), to identify major policy questions and research needs, and to consider scientific and technical issues for continued board attention.

It is important to note that in a single half-day workshop it was not possible to explore the myriad issues and research efforts in progress relating to standards, assessment, student learning, and performance in depth. The intent of the workshop, rather, was to initiate a discussion with a diverse group of scholars and practitioners and develop a sharpened focus on the types of problems and opportunities posed by the new Title I legislation. The board believes that asking questions and developing even better questions are critically important components to sound policy analysis.

Participants included board members and staff, officials of the U.S. Department of Education, state education agency officials, a representative of one of the major test publishers specializing in state and district standardized achievement tests, congressional staff, and others with expertise and interest in Title I (see the Appendix to this report). The agenda (also in the Appendix) included short presentations on specific aspects of Title I assessment, followed by group discussions. This bulletin summarizes the proceedings of the workshop. The opinions voiced at the workshop, which are reported here, do not represent official recommendations of the Board on Testing and Assessment, the National Research Council, or the National Academy of Sciences.

The work of the board is made possible through generous grants from the U.S. Departments of Defense, Education, and Labor. The continued interest and encouragement offered by Steve Sellman, Jane Arabian, Emerson Elliott, Jeanne Griffith, Gary Phillips, Sharif Shakrani, Alan Ginsburg, Valena Plisko, Elois Scott, Raymond Uhalde, Robert Litman, and Donna Dye are very much appreciated.

We wish to express special thanks to board member Richard Elmore for his help in conceiving the workshop described in this report and for agreeing to give up another Saturday in service to the board. Many thanks also to board members Sylvia Johnson, Luis Laosa, and William Taylor, and to staff members Ella Cleveland and Patricia Morison for helpful comments on earlier drafts. Board director Michael Feuer and consultant Nancy Kober were responsible for drafting the workshop summary.

Finally, special thanks to Christine McShane for her editorial skill and patience, and to Eugenia Grohman for her management of the review process. Holly Wells provided excellent administrative support, as usual, and was ably assisted by Adrienne Crawford.

Richard J. Shavelson, Chair
Board on Testing and Assessment

Title I Testing and Assessment

Challenging Standards for Disadvantaged Children

Title I Testing and Assessment

Challenging Standards for Disadvantaged Children

Congress and the Clinton administration had several goals in mind when they revamped Title I testing provisions in the 1994 reauthorization of the Elementary and Secondary Education Act (renamed the Improving America's Schools Act). The new law (P.L. 103-382) shifts important responsibilities from the federal to the state levels, uses Title I as a catalyst for general school reform while promoting challenging learning expectations for students in poor schools, reduces federal reporting requirements, and focuses greater attention on results and relatively less on procedural compliance.

The new law directs the states, by 1997-98, to develop or adopt challenging content standards for Title I, specifying what children should know and be able to do in reading and mathematics, at a minimum. States also must develop or adopt challenging student performance standards that describe proficient, advanced, and partially proficient levels of performance. States that have already developed standards under *Goals 2000* or another process will be able to use those standards for Title I. Those that have not will set standards for Title I students that are consistent with their learning expectations for all children.

By school year 2000-2001, states must develop or adopt high-quality yearly assessments for Title I, in reading and math at a minimum, that are aligned with their content and performance standards. These assessments will be administered at some time during grades 3 through 5, 6 through 9, and 10 through 12. The law lays out several other criteria for new Title I assessments. They should:

- be used only for purposes for which they are valid and reliable,
- be consistent with nationally recognized professional and technical standards,

- to the extent practicable, assess limited-English-proficient (LEP) students in the language and form most likely to yield accurate and reliable information on what such students know and can do,
- make reasonable adaptations for students with diverse learning needs, and
- produce results that can be disaggregated by gender, major racial and ethnic group, English proficiency status, migrant status, disability and nondisability status, and economic disadvantage.

States have another important responsibility under the new law: to decide what constitutes adequate yearly progress in Title I for schools and school districts, using results from the new assessments and other measures. Schools that do not show adequate progress in helping students reach performance standards will be targeted for improvement, will become eligible to receive technical assistance, and eventually will be subject to corrective actions. Similarly, school districts will have to demonstrate adequate progress to their respective state education authorities. Assessment results also will be used to designate "distinguished schools" that are making exceptional progress. Clearly, the law intends for new types of tests and assessments to be used for "high stakes" accountability (Office of Technology Assessment, 1992; Feuer and Kober, 1995; Cronbach, 1995).

The law allows states to propose a set of "transitional" assessments and measures of progress to be used while states are developing new assessment systems. In addition, states will develop benchmarks, timetables, and reporting schedules for completing and field testing their new Title I assessment systems during the next five years.

FRAMING QUESTIONS

Mary Jean LeTendre, Title I director in the U.S. Department of Education, opened the workshop dialogue by listing several Title I assessment issues that could benefit from additional advice, clarification, research, and technical support. Workshop participants discussed these issues and amplified her list.

Transition Policies

The most urgent decisions facing state and federal policy makers pertain to procedures for the transition period, which could last up to six years. (In certain instances, states

may take up to school year 2001-2002 to field-test new assessments.) The department wants to encourage states to use the transition period to develop genuine innovations in their testing programs, although preliminary evidence suggests how challenging it is for most states to move beyond traditional methods. Ideally, procedures developed during the transition period will form a logical connection to the state's long-term standards and assessment goals—for example, by using similar kinds of core measures or differentiated proficiency levels. The transition system should also be rigorous enough to enable the use of test data in a strong system of accountability; this means that the transitional tests will need to provide reliable and comparable information, so that schools and school systems are judged fairly.

The discussion highlighted several transition issues for which further technical advice and information could be useful. For example, how can the Department of Education judge the appropriateness of state plans and benchmarks for developing new assessment systems? How can the department determine whether proposed transition methods for determining adequate progress of schools are based on accurate information, as the law requires? States need information about the wide range of appropriate measures and assessments that could be used during the transition to determine student and school performance; they will also need technical assistance in choosing from new and/or available options for setting student proficiency levels.

An important determinant of state decision making during the transition period will be financial: school budgets will impose limitations on states, forcing them to exercise prudence in the choice between investing in transition procedures and investing in long-term testing systems.

Assessment of Young Children

Under the new law, formal Title I assessments will not begin until grades 3 through 5. Yet nearly half the children served in Title I—49 percent in school year 1992-93—are in grades prekindergarten through 3. A key question, therefore, is whether there will be a need for a special assessment system targeted to children in grades preK through 2, to ensure that they are being taught the same knowledge and skills as other children and are acquiring a solid foundation for future learning. Growing awareness of the critical importance of early childhood learning for later achievement leads to a widely felt need to hold schools accountable for quality education in the early grades and to establish formal links between early instruction and later student performance. But the states will need especially rigorous technical advice on appropriate strategies for assessing younger chil-

dren in Title I, given the special needs of young children with respect to test design, administration, and reporting of results to teachers and parents (see, for example, Shepard, 1991).

Adequate Progress

The law directs states to define adequate progress for schools and districts in a way that results in continuous and substantial yearly improvement, sufficient to enable all Title I students to meet state proficiency standards. The Department of Education is seeking input on multiple definitions and various models for adequate progress, and there is a growing awareness of the importance of the continuous improvement concept. Kentucky, for example, has established a 20-year period for achieving its state improvement goals, with benchmarks that assume steady growth every 2 years. Although the Kentucky accountability system has recently been criticized (reference to come), its experience in a long-term approach to progress is worthy of close attention.

Measurement Issues

With the new Title I assessment provisions, the nation has the opportunity "to make a dramatic statement about where we need to go," said Thomas Payzant, who was Assistant Secretary for Elementary and Secondary Education in the U.S. Department of Education and has now become schools chancellor in Boston. The problem, he said, is that many current testing tools influence schools to adopt behaviors contrary to the goals of teaching challenging content and skills to Title I children.

The new law encourages states and school districts to adopt assessments that are both rigorous and innovative. But along the way, many technical measurement issues must be addressed. What, for example, do validity and reliability mean in the context of various kinds of performance-based assessment? How can the Department of Education provide guidance on technical testing issues while avoiding the specificity and rigidity that characterized prior Title I testing guidelines?

Multiple Measures

According to the new law, state assessment systems should involve multiple up-to-date measures of student performance, including measures that assess higher order thinking skills and understanding—a provision that encourages performance-based assessments and discourages overreliance on a single test score. But the concept of multiple measures is new in Title I testing. What might these measures include? How can they be applied simultaneously? How should states and schools balance information from dif-

ferent measures, particularly if there is no consensus? Eva Baker of the University of California at Los Angeles cautioned that, without some stipulations, schools could emphasize the measure that looks the most positive. The history of test-based accountability holds important lessons for the states as they develop new Title I standards and assessment systems (Koretz, et al., 1991; see also the report of the select committee on the California Learning Assessment System in Cronbach, 1995).

Limited-English-Proficient (LEP) Children

The Title I law directs states to assess LEP children, to the extent practicable, in the language and form most likely to yield accurate and reliable information on what such students know and can do. The intent of the law seems to be to urge the development of accommodations in testing, including (but not limited to) providing tests in students' home languages. Other accommodations typically mentioned by advocates of increased inclusion of special populations in large-scale assessments are increased testing time and provision of glossaries to aid students whose native language is not the language of the test. Implementing this provision poses some formidable challenges. Existing assessments in other languages are not always of high quality, nor are they necessarily comparable in difficulty or scoring to their companion tests in English. In most languages, assessments are not yet available. Recognizing this, the Title I law charges the Department of Education with helping states identify appropriate assessments in needed languages. What role should the federal government play in encouraging development of assessments in different languages? And how can schools measure progress of LEP students while better multilingual assessments are being developed?

Disaggregated Data

The requirement for disaggregated data also raises questions. How can states create assessments that lend themselves to disaggregation but do not entail excessive testing burden? How should schools and districts report disaggregated data? In small schools, would disaggregation of results create potential breaches of privacy (if, for example, there are very few students who fall into any of the required classifications)?

STATE PRACTITIONER PERSPECTIVES

Joseph Johnson of the Texas Education Agency launched a discussion of critical issues facing states as they try to fashion an effective and workable Title I assessment system. It

is a process, participants agreed, that requires continual monitoring and willingness to change components that do not work as planned. Johnson made four major suggestions, drawn from his state's experience in shifting from an accountability system based on process to one based on results.

Politics of Standards

Establishing standards for student learning can become a highly controversial process. Some people may see state standards as a threat to local control. Others may confuse standards-based reforms with the controversial issue of outcome-based education. And it takes courage, Johnson said, to set rigorous standards that exceed the proficiency levels that many students now exhibit. He recommended that states and school districts develop strategies to communicate to the public that high expectations are reasonable and appropriate.

Clear and Understandable Assessments

State assessment systems need to be understood by parents, teachers, students, and the public. One reason why prior Title I testing procedures did not drive better instructional practices, for example, was that most teachers, principals, and parents did not really know what a normal-curve equivalent score and other similarly technical language signified. (Texas still has some problems with the clarity of its assessment system. The state uses scaled scores, in which 70 is a passing score, which can be misleading for parents who expect a score of 100 to mean that a student answered all questions correctly.) Baker raised the possibility of a trade-off between technical purity of tests and the degree to which their content and results are credible and understandable to the public.

Accuracy and Efficiency

State Title I assessments must give accurate information about student performance, including capturing gradations of performance below the proficient level. But accuracy is not the sole consideration. A system that provides the most accurate information might entail too much testing time, or cost too much money, or might be impractical in other ways. Furthermore, states should recognize that performance-based assessments may not be as accurate as traditional kinds of assessments in measuring certain skills and knowledge areas.

Fairness

States also must consider fairness issues in developing new standards and assessments. Johnson gave the example of an English-language reading test that was translated literally into Spanish for Spanish-speaking students but was neither fair nor accurate. It is important to note that lower scores among specific groups of students do not automatically mean that a test is unfair, especially if the questions answered incorrectly pertain to skills and understanding that all children should have but that the lower-scoring children were not taught.

Driving Desired Behaviors

New Title I assessment systems should encourage sound instruction and appropriate behaviors, Johnson said. For this to occur, teachers, principals, parents, and students must see the relationship between student assessment results and classroom practices. Teachers must understand what kinds of instructional and curricular changes will produce higher performance. Baker cautioned that "it is a fallacy to think teachers know how to do this." Teachers, principals, and others will need training and guidance about how to use assessment results to modify instruction and how to interpret disaggregated data to improve learning for particular groups of students.

Unintended consequences of high-stakes testing programs need to be anticipated, to the extent possible. In Texas, for example, the initial assessment system focused on the percentage of children who passed, which inadvertently encouraged schools to concentrate maximum effort on those children who fell just below the passing score and to make only token efforts with those who were substantially below. William Taylor, a member of the Board on Testing and Assessment, suggested that states might take care of this problem by carefully defining the partially proficient performance level.

Districts and states can also influence behaviors through corrective actions for schools that are not making progress. The challenge is knowing when and how to intervene.

MEASUREMENT ISSUES

Leading a discussion of measurement issues, Eva Baker stressed the need for broad experimentation with different assessment approaches.

Quality of Performance Assessments

Baker identified five questions that states and school districts might ask when judging the quality and suitability of performance assessments for Title I:

- Are the assessments of high technical quality? (Most existing ones are not, Baker said.)
- Do they measure progress toward the chosen standards or curricular goals?
- Do they guide instruction in productive directions?
- Are they fair and will they raise the standards for all children, including Title I children?
- Are they sensitive to reform? In other words, Baker explained, once the novelty wears off, how long will it take to improve them to the point that they meet high technical and professional measurement criteria?

Measurement Challenges

Baker's comments prompted discussion about other measurement challenges associated with developing Title I assessments. One such challenge is determining how generic the purposes of the assessment will be. Often there is pressure to come up with a one-size-fits-all assessment, Baker said, a strategy that has the added benefit of sending a coherent message.

Baker advised states to decide up front whether the primary purpose of the assessment will be instructional improvement or public accountability, because this will affect the design process.

Another challenge is defining and validating proficiency levels—in particular, defining the partially proficient category in a way that reflects how children build knowledge and progress toward higher achievement levels. Often children show different rates of improvement in different subjects or have spurts of intellectual growth rather than steady advancement.

Crafting performance assessment tasks that effectively sample a domain of knowledge is another measurement challenge. Performance assessments vest much significance in a single task. But what if the task chosen is something that the test taker was never taught? What inferences can be made from the results?

Other measurement questions will require technical guidance—for example, designing and scoring performance tasks, developing new models for appropriate use, phasing in

new assessments, determining appropriate combinations of data from multiple measures, and developing measures for grades K through 3.

Reporting Challenges

Another set of questions relates to the reporting of assessment data. Against what standards will data be reported? Will the data be reported uniformly for every school? Will the reporting system allow for mitigating circumstances or reporting exclusions—for example, students enrolled in a particular school for just one day? What kinds of comparisons should be made among groups or school sites? In what format should data be reported?

Disaggregated Data

Workshop discussion suggested that the directive for disaggregated data will be especially complicated from a measurement standpoint. The Title I law seems intended to exempt states from the disaggregation requirement in instances in which the data would not be statistically sound. But the law does not specify how small a group is too small or how individual student scores could be safeguarded in small-numbers situations. The new law mandates disaggregation for at least two main reasons: to enable statistically sound descriptions of performance of particular groups of students (e.g., limited-English proficient, disadvantaged, etc.), and to enable policy makers and the public to draw inferences about the extent to which LEP and other students served by Title I are making adequate yearly progress toward meeting the performance standards. These purposes are consistent with the law's general concern for equity in standards and testing and for the concept that all students can perform to high standards. The law also clearly intends to encourage alternative assessments, i.e., to reduce the reliance on standardized instruments; the problem is that some states seem to believe that the requirement for disaggregation can be met only with commercially available, standardized assessments.[1]

Board member and workshop chair Richard Elmore suggested that states, test publishers, researchers, and others mount small trials exploring various solutions to these and other common measurement problems. Results from these experiments could provide a base of information to help states make decisions.

---

[1] Technical questions that surround the disaggregation requirement were mentioned during the workshop, but not discussed or analyzed in detail. Among the issues that warrant further attention are: sample size requirements to yield valid inferences about student subgroups, effects of small sample sizes on the confidentiality of individual student data, and reporting of data at school, district, or state levels depending on the uses to which the requested data will be put. These questions may be addressed in the future work of the board.

TEST PUBLISHER PERSPECTIVES

A presentation by Tom Williams of CTB-McGraw Hill anchored a discussion of test publishing and commercial development issues. Williams mentioned several factors that he felt will affect the success of new standards and assessments.

Clarity of Standards

Standards should be clear and understandable. Some existing state standards do not meet this test. Even if consensus develops around a set of standards, said Williams, the standards will not remain stable because different people can redefine the same words in different ways and standards can be modified by legislation.

Standards as "Hard Currency"

Standards should reflect something meaningful in the real world, Williams said. For example, there should be demonstrable rewards or incentives—in terms of success in the workplace or higher education—for students who achieve the standards.

Technical Quality

Test publishers claim to have made substantial progress in improving performance assessment and in attaining higher levels of quality in these assessments, within certain limits. But even if the commercial publishers create new performance assessments that are reasonably valid and cost effective, formidable barriers remain before reaching acceptable levels of generalizability.

Willingness to Invest

The success of new standards and assessments will depend on the willingness of many sectors to invest time and money: time is needed to field test new assessments, allow teachers to practice using and scoring them, and provide staff development, and money is needed to develop or purchase assessments, conduct research and validation studies, and improve their quality.

Competitive Market

Williams argued that a competitive marketplace—"in which the clients are free from repressive mandates and from arrogance of outsiders and others who think they know

best"—is more likely to produce successful assessments than a restrictive one. Test publishers and their clients (school districts and states) need freedom to experiment and use marketplace resources, he said, and freedom from political interference. "Don't restrict [the market] to a small number of vendors," he cautioned. "Let competition work."

Williams further urged that state and local educational agencies be given freedom to design tests "without prejudice about format." If they decide that a multiple-choice format works best for measuring a particular skill, then they should be able to use it. If states and locals really have the responsibility, then their judgment must be respected."

Consumer Awareness

Williams suggested that the public and private sectors share information and undertake stronger partnership efforts to improve consumer awareness of standards and assessment issues.

ACCOUNTABILITY ISSUES

Helen Ladd of the Brookings Institution and Duke University drew from her research on performance-based accountability systems in the state of South Carolina and the Dallas, Texas, Independent School District, and from other policy analysis research to lead a discussion of school governance and accountability.

Unit of Accountability

Based on research showing that teachers alone do not have sufficient power to change school culture, Ladd recommended that the *school* be the unit of focus in performance-based accountability systems and advised against focusing rewards or sanctions on individual teachers.

Weighing SES Factors

A difficult issue is whether accountability systems should make adjustments for factors that are typically considered to fall outside the control of the school, such as socioeconomic background. In other words, if the socioeconomic background of students is known to be a correlate of their achievement, then should this information be integrated into a fair system of accountability that holds schools responsible for progress in student achievement?

The accountability programs in Dallas and South Carolina consider these factors but use different approaches. Both Dallas and South Carolina base their accountability program on test scores for every student every year, and both use sophisticated (multiple regression-based) methodology to match student data longitudinally and produce individual and aggregate school gain scores. Dallas uses mean test score gains in multiple subjects; South Carolina uses median scores. Dallas also uses multiple tests and measures—including a nationally normed test, the Texas state assessment, and more—so that teachers do not teach to one test.

The two sites use different approaches to calculate student gain scores, and both have modified their systems over time. Originally the South Carolina program, established in 1984, calculated gain scores by comparing a student's actual test score in a given year with the predicted score for that year. The predicted score was calculated solely on the basis of the previous year's test score, without taking into account socioeconomic status (SES), an approach that the state felt would be the fairest system. But the state soon realized that the schools with the top aggregated gain scores were those with the highest SES characteristics. South Carolina changed its approach by clustering schools into five categories based on SES characteristics (but not racial and ethnic characteristics) and comparing only those schools within the same categories. The intent is to recognize that SES affects learning and that different schools face different challenges. This new approach is not without problems, Ladd said. Schools care what category they are in because the cut points for rewards and sanctions are different in different categories.

Dallas takes into account race, gender, income, and other factors in predicting student gain scores. Although the Dallas methodology yields results that do not show pronounced SES difference in school outcomes, the trade-off is that it sends a signal of lower expectations for learning among certain groups of children.

This issue—accounting for what economists call exogenous determinants of achievement—raises fundamental policy dilemmas. Does the assumption that some variables are out of the control of schools essentially let them off the hook in terms of accountability? Or does preventing the inclusion of these variables in an accountability system lead to an unfair system of judgment? Some educators argue that schools do indeed have control over the variables necessary to ensure the success of their students. Joseph Johnson, for example, pointed out that many low-SES schools defy conventional wisdom and achieve at levels comparable to some of the wealthiest schools; among them are the 81 schools in Texas, including 25 schoolwide projects, in which 65 percent or more of Title I students passed all sections of the state assessment. None of the

schools is from Dallas, said Johnson, attributing this in part to his belief that the Dallas assessment approach does not articulate high expectations for all students.

The Dallas assessment results have been used to make decisions about reassigning and discharging principals, Ladd said, a strategy that may eventually pay off with instructional improvements. This could have a downside, however, according to Lloyd Bond of the University of North Carolina at Greensboro, if it makes it harder to recruit inner-city teachers and principals: "If accountability goes to principals and teachers, no one will want to be the principal of an inner-city school." Ladd disagreed, asserting that teachers and principals feel they are being treated more fairly when SES conditions are taken into consideration.

It is important to position this ideological and methodological debate in the apparent intentions of the law. For example, according to William Taylor, neither the Dallas nor the South Carolina model would be appropriate for judging the adequate progress of schools under Title I, because both systems presume slower progress for lower-SES students and apply factors to compensate. The new Title I law, by contrast, expects faster progress among these children, with the goal that they will eventually catch up with their peers. Bond concurred, saying that both the South Carolina and Dallas systems presume that low-achieving children will never close the gap. The technical and policy challenge, then, is devising a fair system of accountability that does not send unintended signals of differential expectations for students of different backgrounds.

Public Understanding

Helen Ladd and Joseph Johnson agreed on a key point, namely that the Dallas system is complicated and is not well understood even among people in the district. Ladd felt that the creation of a Dallas accountability task force, comprised of lay people who make policy decisions, has helped, but that Dallas educators still do not understand fully what instructional changes are required to improve assessment results.

Eva Baker said that, if the public trusts the people making the assessment decisions, then the assessments can be more complex. Johnson noted that principals and teachers have to see the relationships between assessment results and classroom practice in order for the assessments to have real meaning. The lack of understanding in Dallas has created fear and frustration, he added—fear because principals know that their job is on the line, but frustration because they may not know what they need to do to make student scores better. The consequence is that educators do not always focus their energies in constructive directions.

ISSUES FOR FURTHER CONSIDERATION

The workshop discussion pointed to several issues to be considered by federal and state policy makers in developing new Title I standards and assessments. Several of these issues could benefit from technical advice and support from the Board on Testing and Assessment.

- **New Title I assessment systems should be understandable to those who must implement them and clear enough to attract public support.**

"There is a scary activism growing out there," said Eva Baker, and it is based on dangerous misinformation about what standards do and do not do. Some members of the public and the media erroneously believe that standards-based assessment means avoiding hard statistical measures, that standards are the same as outcome-based education, and that national standards are being forced on states and school districts. It is important to be very precise in explaining standards to the public and avoid such charged terms as outcomes and higher-order thinking skills.

- **New assessment systems must balance technical quality and measurement issues with such practical issues as cost, transparency, and ease of implementation.**

Policy makers will have to make difficult trade-offs in the months ahead. An assessment system that is a technical wonder may be too complicated to understand or too expensive to pilot-test and implement.

- **There is a need for additional research, demonstration, and field-testing on the technical aspects of performance assessment.**

Performance assessment is so new that many measurement issues have not been resolved; traditional models of test development and measurement theory do not necessarily apply (see Linn, 1994; Shavelson, et al., 1993).

- **New accountability systems should be fair to schools that serve low-SES populations but should not signal lower expectations for any groups of students.**

This will be a difficult balance to achieve. Richard Elmore suggested that states define adequate progress as more than a single cutoff score, perhaps reflecting dimensions of time, need, and other relevant factors.

- **New assessments should reward the right kinds of instructional practices and school behaviors.**

As Thomas Payzant and other participants emphasized, standards and assessment are part of larger reforms in instructional practice. Aligning curriculum, standards, and assessment is a primary goal of the new Title I.

- **Policies for assessing progress during the transition period should flow logically into long-term state assessment goals for Title I.**

States could benefit from diverse models of transition policies that lay the groundwork for a better long-term assessment system for Title I. For example, although there is no strict requirement to disaggregate during the transition period, it still may be useful for states to use some kind of disaggregation. It was suggested that the Board on Testing and Assessment help develop models for achieving disaggregation with performance-based assessments, and provide advice on the kinds of disaggregated data necessary to form a baseline in a period of great change. Without this type of guidance, states might feel compelled to use conventional assessments for the short term.

- **Federal and state agencies, advisory groups, and the private sector should make available a wide range of models illustrating many creative and flexible ways to implement new Title I assessments.**

States need information about a range of options for addressing the goals of the new Title I assessment provisions. The Board on Testing and Assessment or other research groups could provide information to states, schools, parents, and others, describing and discussing various innovative models for standards and assessment.

- **Policy makers, educators, the general public, and the private sector must view the development of standards and assessments as an ongoing process that will not end in 2001.**

Even after new Title I standards and assessments are implemented, there will be a continuing need for research, monitoring, revision, and improvement. This is particularly true of performance assessments and other innovative measures.

REFERENCES

Cronbach, Lee J.

1995 *A Valedictory: Reflections on 60 Years in Educational Testing.* Board Bulletin. Includes the report of the select committee on the California Learning Assessment System. Board on Testing and Assessment, National Research Council. Washington, D.C.: National Academy Press.

Feuer, Michael J., and Nancy Kober

1995 *Anticipating Goals 2000: Standards, Assessment, and Public Policy.* Board Bulletin. Board on Testing and Assessment, National Research Council. Washington, D.C.: National Academy Press.

Koretz, Daniel, Robert Linn, Stephen Dunbar, and Lorrie Shepard

1991 The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Ill., April 1991.

Linn, Robert

1994 Performance assessment: Policy promises and technical measurement. *Educational Researcher* 23(9):4-14

Office of Technology Assessment, U.S. Congress

1992 *Testing in American Schools: Asking the Right Questions.* Washington, D.C.: U.S. Government Printing Office

Shavelson, Richard, X. Gao, and Gail Baxter

1993 Sampling variability of performance assessments. *Journal of Educational Measurement* 30(3):215-232

Shepard, Lorrie

1991 Readiness testing in local school districts: An analysis of backdoor policies. Pp. 159-180 in Susan Fuhrman and Betty Malen, eds., *The Politics of Curriculum and Testing.* New York: Falmer Press.

APPENDIX

WORKSHOP AGENDA AND PARTICIPANTS

AGENDA

BOARD ON TESTING AND ASSESSMENT

Green Building, Room 110
National Research Council
2001 Wisconsin Avenue
Washington, DC

April 1, 1995

Richard F. Elmore, Chair

12:45 pm	Welcome and Introductions
1:00	Framing Questions (US ED staff)
	♦ What is the prospect for the development and implementation of performance measures that meet the new requirements of Title I?
	♦ What criteria could be used by states to identify schools and LEAs for program improvement during the time they are using transitional assessments?
	♦ What level of performance would be expected during the time states are using transitional assessments?
1:15	A State Practitioner's Perspective Joseph Johnson, Texas Education Agency
2:00	Measurement Issues Eva Baker, UCLA

2:45	A Test Publisher's Perspective Tom Williams, CTB-McGraw Hill
3:30	General Issues in Educational Accountability and Governance Helen Ladd, Brookings Institution
4:15	Next Steps Developing and communicating effective implementation options Richard Elmore, BOTA
5:00	Adjourn

PARTICIPANTS

EVA BAKER, Graduate School of Education, University of California at Los Angeles

CHARLES BARONE, Office of Senator Paul Simon

JOANNE BOGART, Planning and Evaluation Service, U.S. Department of Education

LLOYD BOND, School of Education, University of North Carolina, Greensboro

JOSEPH CONATY, Office of Educational Research and Improvement, U.S. Department of Education

EMERSON ELLIOTT, National Center for Education Statistics, U.S. Department of Education

RICHARD ELMORE, Graduate School of Education, Harvard University

MICHAEL FEUER, National Research Council, Washington, D.C.

EUGENE GARCIA, Bilingual Education and Minority Language Affairs, U.S. Department of Education

ELLEN GUINEY, Labor and Human Resources Committee, U.S. Senate

JOSEPH JOHNSON, Texas Education Agency, Austin

SYLVIA JOHNSON, School of Education, Howard University, Washington, D.C.

NANCY KOBER, Charlottesville, Va.

HELEN LADD, Brookings Institution, Washington, D.C.

MARY JEAN LeTENDRE, Title I Programs, U.S. Department of Education

WILLIAM LOBOSCO, Title I Programs, U.S. Department of Education

PHYLLIS McCLURE, Washington, D.C.

PATRICIA MORISON, National Research Council, Washington, D.C.

FREDERIC MOSHER, Carnegie Corporation of New York, New York

THOMAS PAYZANT, Elementary and Secondary Education, U.S. Department of Education

VALENA PLISKO, Elementary and Secondary Education, U.S. Department of Education

GRACE ROSS, Title I Programs, U.S. Department of Education

LISA ROSS, Labor and Human Resources Committee, U.S. Senate

ELOIS SCOTT, Planning and Evaluation Service, U.S. Department of Education

SHARIF SHAKRANI, National Center for Education Statistics, U.S. Department of Education

WILLIAM TAYLOR, Washington, D.C.

ALEXANDRA WIGDOR, National Research Council, Washington, D.C.

SUSAN WILHELM, Title I Programs, U.S. Department of Education

TOM WILLIAMS, CTB-McGraw Hill, Monterey, Calif.

About the Board on Testing and Assessment

The Board on Testing and Assessment was established in 1993 with support from the United States Departments of Defense, Education, and Labor. Its principal objectives are to aid policy makers in the clarification of the purposes of testing and assessment and to help them evaluate the use of tests, alternative assessments, and other indicators commonly used as tools of public policy. The board brings to bear the knowledge and tools of the social and behavioral sciences and provides an analytical base for the examination of difficult issues in measurement and evaluation as they emerge in education, the workplace, and other settings. The board is a long-term activity of the National Research Council designed to be responsive to evolving challenges that face schooling, work, and the measurement of human competencies. Some specific functions of the board include analyzing implications of the science of testing and assessment, providing a neutral forum for parties to address issues which to discuss the effects of planned testing and assessment policies, helping government agencies coordinate their policies, and conducting in-depth studies of technical and policy problems in testing and assessment.

Members 1995-1996

Richard J. Shavelson (*Chair*), Stanford University
Lauress J. Bass (*Vice Chair*), Georgetown University
Robert L. Linn (*Vice Chair*), University of Colorado, Boulder
Richard C. Atkinson, University of California, Oakland
Jeanne G. Barnes, Potomac Electric Power Company
David C. Berliner, Arizona State University, Tempe
Paul J. Black, King's College London
Richard F. Elmore, Harvard University
Arthur S. Goldberger, University of Wisconsin, Madison
Edmund W. Gordon, City University of New York
Paul Holland, University of California, Berkeley
Sylvan T. Johnson, Harvard University
Carl F. Kaestle, University of Chicago
Michael W. Kirst, Stanford University
Luis M. Laosa, Educational Testing Service
Renee S. Lerche, Ford Motor Company
Alan M. Leventhal, University of Pittsburgh
James L. Outtz, Outtz and Associates
Paul R. Sackett, University of Minnesota, Minneapolis
Alan H. Schoenfeld, University of California, Berkeley
William L. Taylor, Attorney at Law, Washington, D.C.
Ewart A.C. Thomas, Stanford University
Jack Whalen, Xerox Corporation

BEST COPY AVAILABLE